



<http://esd.lbl.gov/BWC/>

Designing CyberInfrastructure to Support End Science

Deb Agarwal (UCB and LBNL)
Catharine van Ingen (MSFT)
Berkeley Water Center
Microsoft TCI

IndoFlux Meeting, Chennai, India, July 13, 2006

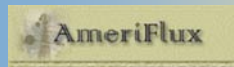


Project Motivation

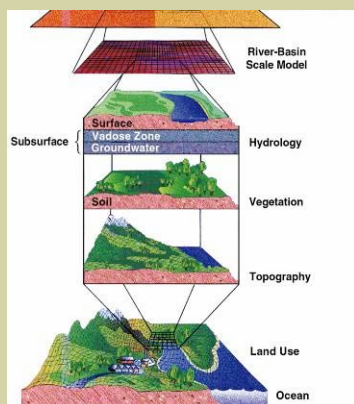
- Data is now being gathered into common data archives
- Data archives provide an opportunity for cross-discipline and cross-site investigations
- Data analysis techniques which worked well on small data sets often do not scale
- Current CS tools have evolved in support of other disciplines – Investigate their ability to facilitate data analysis



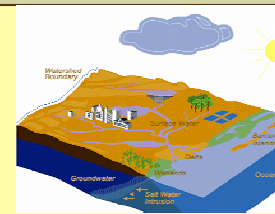
Distributed Data Sets



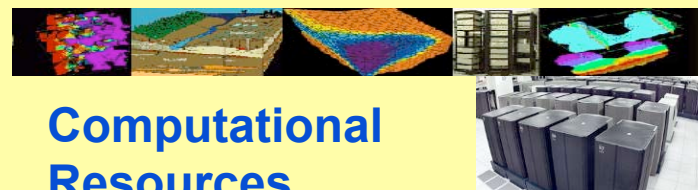
Data Harvesting and Transformations



Data Cleaning, Models, Analysis Tools



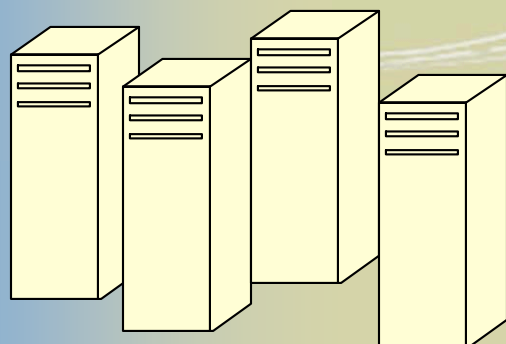
Computational Resources



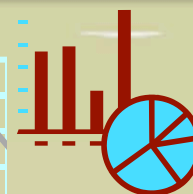
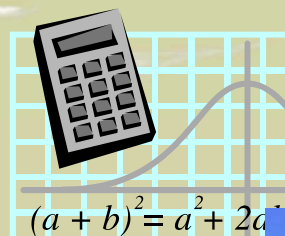
Science Portal

*Building BWC Water
Cyberinfrastructure to
Connect Data,
Resources, and People*





Data Providers:
Host Ameriflux
Climate Data
Statsgo Soils Data
MODIS products



Tools:
Statistical
Graphical



Microsoft

Web Service Interface to Data and Tools

**Web-based
Workbench
access**



**Choose Ameriflux
Area/Transect, Time
Range, Data Type**

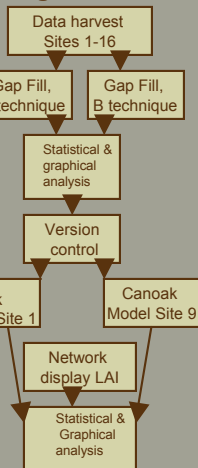


LAI
Temp
Fpar
Veg Index
Surf Refl
NPP
Albedo

**Import other
Datasets**

Climate
Statsgo
MODIS

Design Workflow



Ecology Toolbox

Data
Cleaning Tools

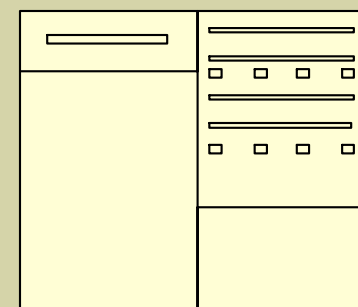
Knowledge Generation Tools

Data Mining
and
Analysis Tools

Modeling Tools

Visualization
Tools

**Compute
Resources**



Carbon Community Workbench



Approach

- Work closely with the end scientists to define, prototype, and test the system
- Provide a solution that leverages both server-based and local desktop/laptop environments
- Leverage commercial tools to the extent possible

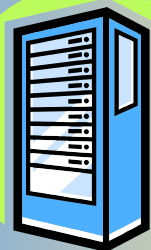


Some Critical Capabilities

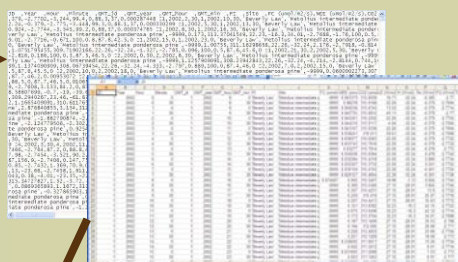
- Support for versioning of data sets
- Work with multiple data sets
- Advanced data selection and plotting capabilities
 - Select data relative to an event
 - Simple calculation across any specified date range
 - Statistical information available
 - Plots - scatter, diurnal, time series, probability density function, tiled, correlation
- Ability to access capabilities from desktop



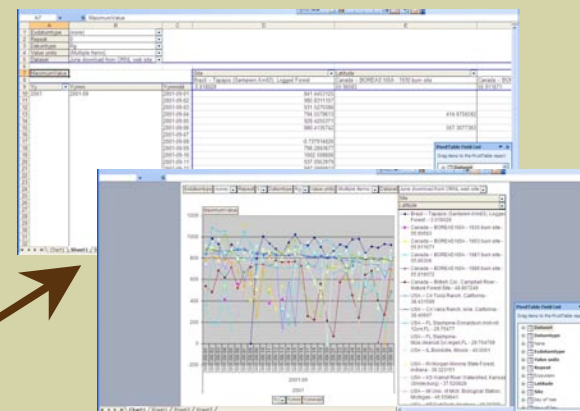
Data Pipeline



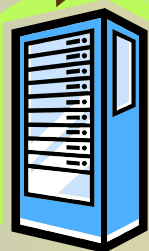
CSV Files



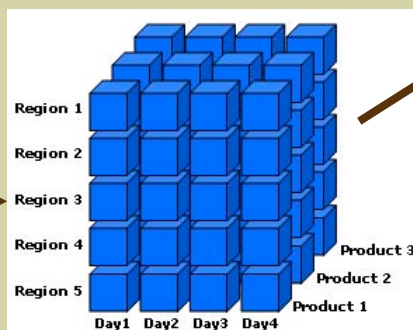
Excel Pivot Table and Chart



ORNL Ameriflux Site



BWC SQL Server Database



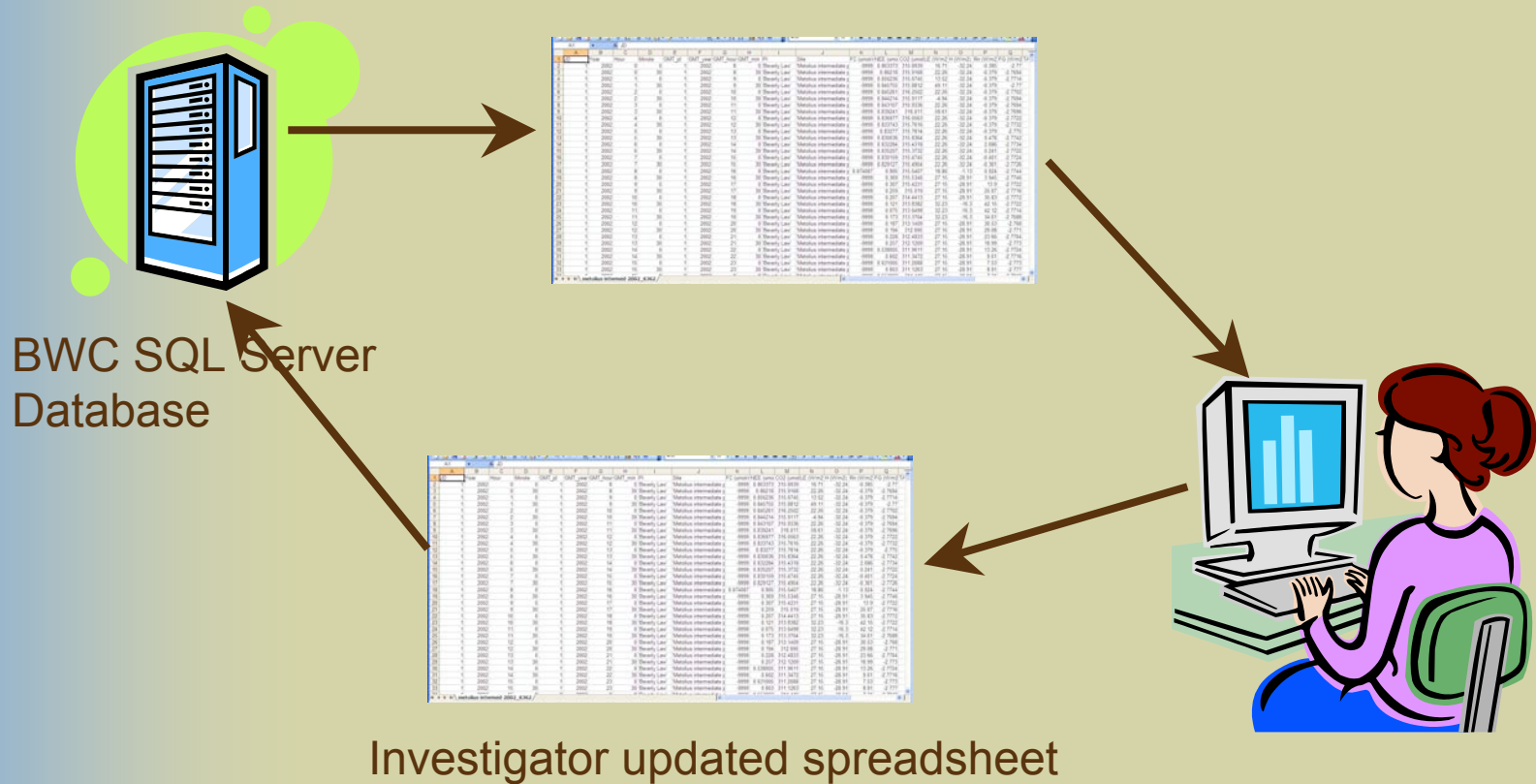
Data Cube





Data Cleaning and Versioning

Excel spreadsheet of current data





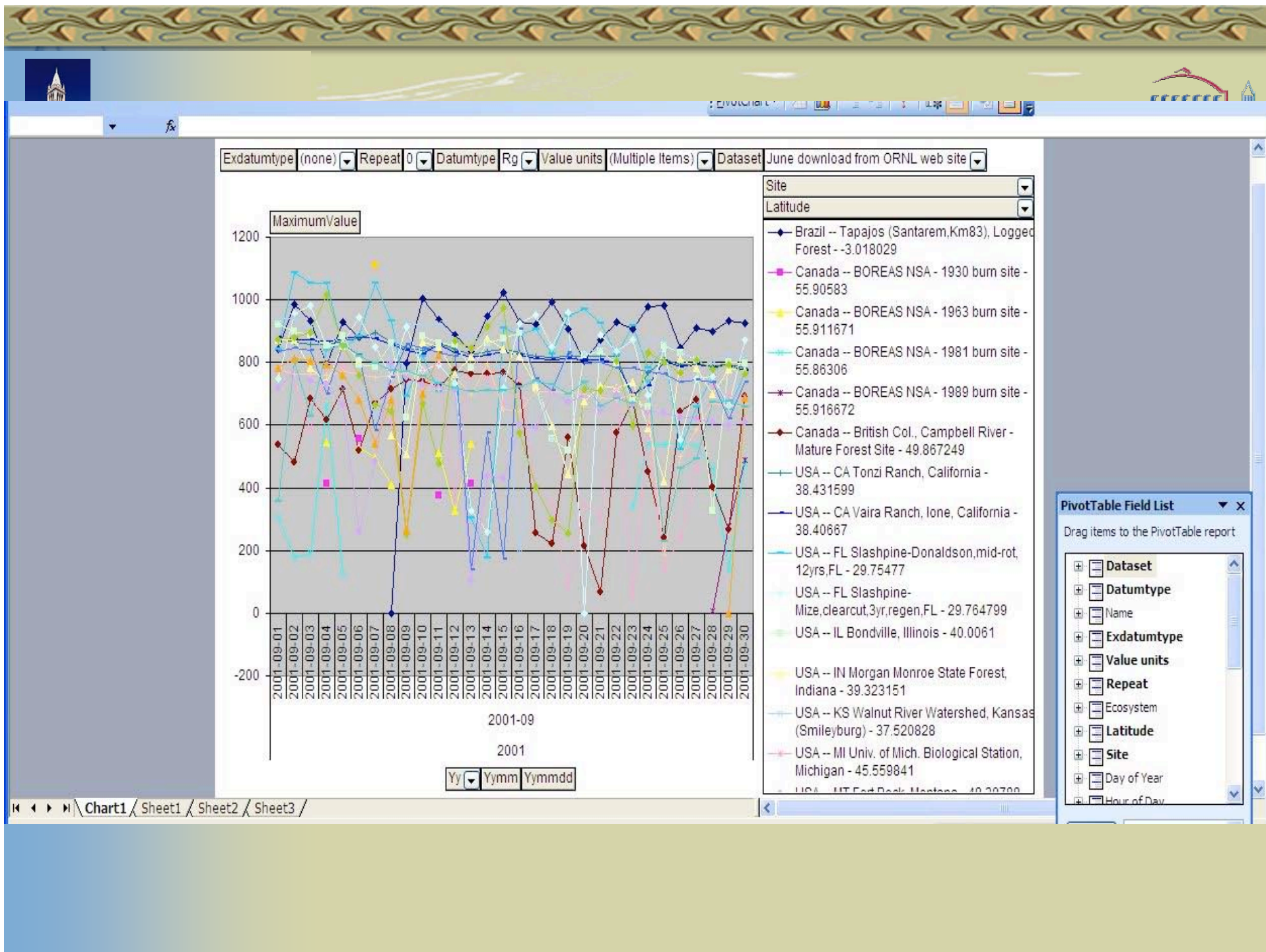
Analysis Services Data Cube

- An organized view of the data
- A multi-dimensional view into the data
- Can integrate multiple data sources
- Define measures and dimensions
 - Measure – a value you want to be able to plot
 - Dimension – An axis you want to be able to use to select data and as axis
- Calculations – define new measures



f MaximumValue						
A	B	C	D	E		
1	Exdatumtype	(none)				
2	Repeat	0				
3	Datumtype	Rg				
4	Value units	(Multiple Items)				
5	Dataset	June download from ORNL web site				
6						
7	MaximumValue		Site	Latitude		
8			Brazil -- Tapajos (Santarem,Km83), Logged Forest	Canada -- BOREAS NSA - 1930 burn site		Canada -- BOF
9	Yy	Yymm	Yymmdd	-3 018029	55 90583	55.911671
10	2001	2001-09	2001-09-01	841.4453125		
11			2001-09-02	985.8311157		
12			2001-09-03	931.5270386		
13			2001-09-04	794.5578613	414.6756592	
14			2001-09-05	928.4255371		
15			2001-09-06	880.4135742	557.3077393	
16			2001-09-07			
17			2001-09-08	-0.737914026		
18			2001-09-09	798.2893677		
19			2001-09-10	1002.508606		
20			2001-09-11	937.0953979		
21			2001-09-12	887.0908813		
22			2001-09-13	834.5820313		
23			2001-09-14	949.5360107		
24			2001-09-15	1021.355408		
25			2001-09-16	929.6103516		
26			2001-09-17	922.444519		
27			2001-09-18	991.7322388		
28			2001-09-19	905.5108032		
29			2001-09-20	804.617981		
30			2001-09-21	873.005188		
31			2001-09-22	930.5013428		
32			2001-09-23	906.8458862		

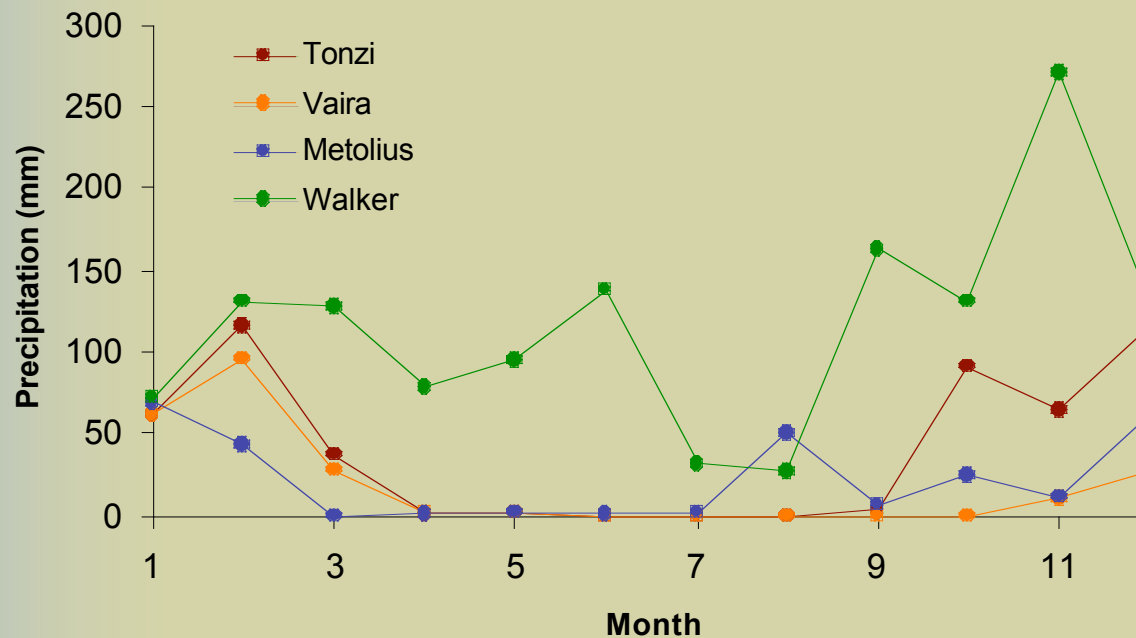
PivotTable Field List	
Drag items to the PivotTable report	
<input type="checkbox"/> Dataset	
<input type="checkbox"/> Datumtype	
<input type="checkbox"/> Name	
<input type="checkbox"/> Exdatumtype	
<input type="checkbox"/> Value units	
<input type="checkbox"/> Repeat	
<input type="checkbox"/> Ecosystem	
<input type="checkbox"/> Latitude	
<input type="checkbox"/> Site	
<input type="checkbox"/> Day of Year	
<input type="checkbox"/> Hour of Day	





Precipitation trends and totals

Precipitation Trends for 2004



Summer precipitation:

Tonzi and Vaira ~ 2% of total

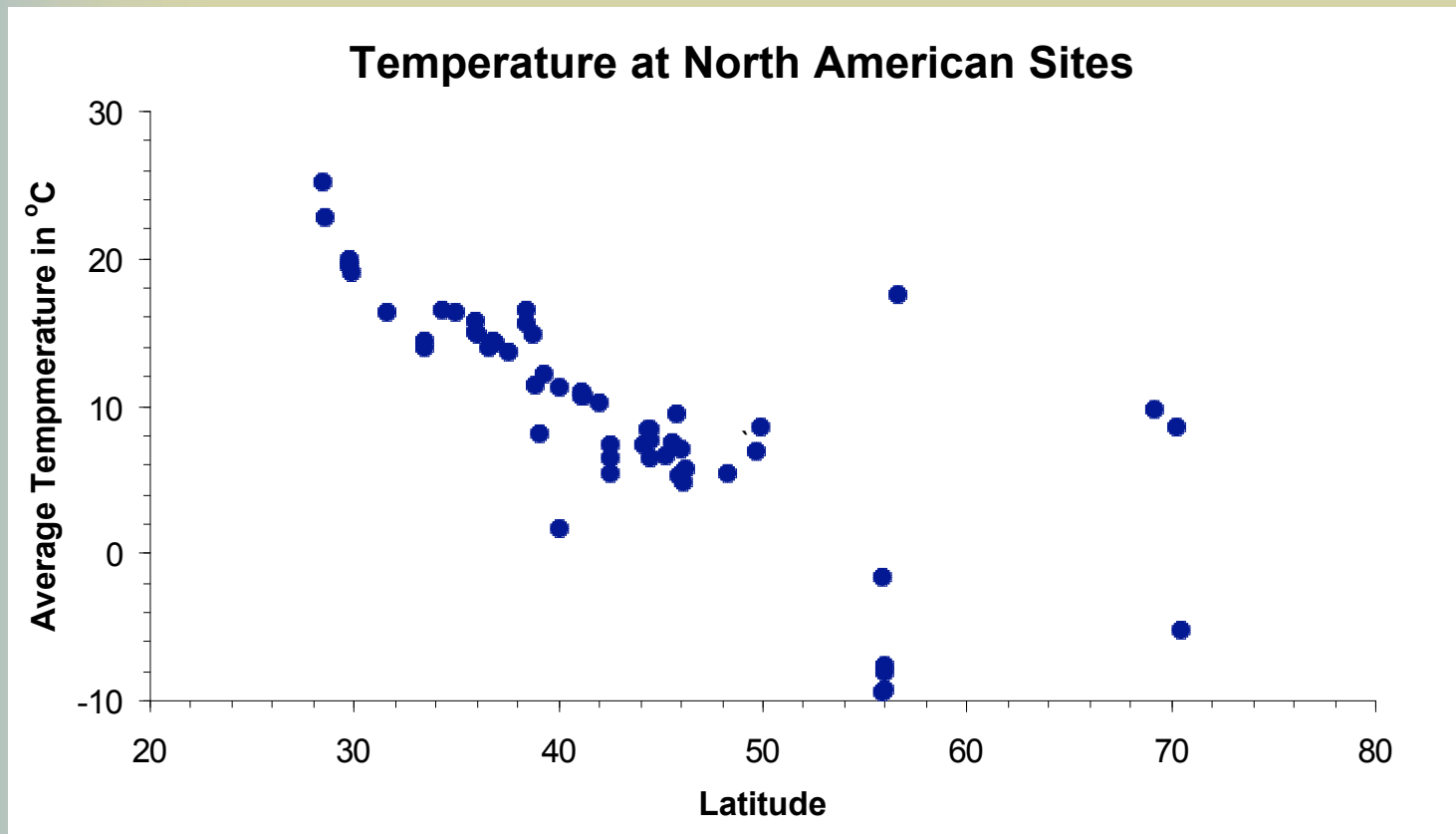
Metolius ~ 24% of total

Walker Branch ~ 40% of total

***Plot created by Gretchen
Miller of UC Berkeley**



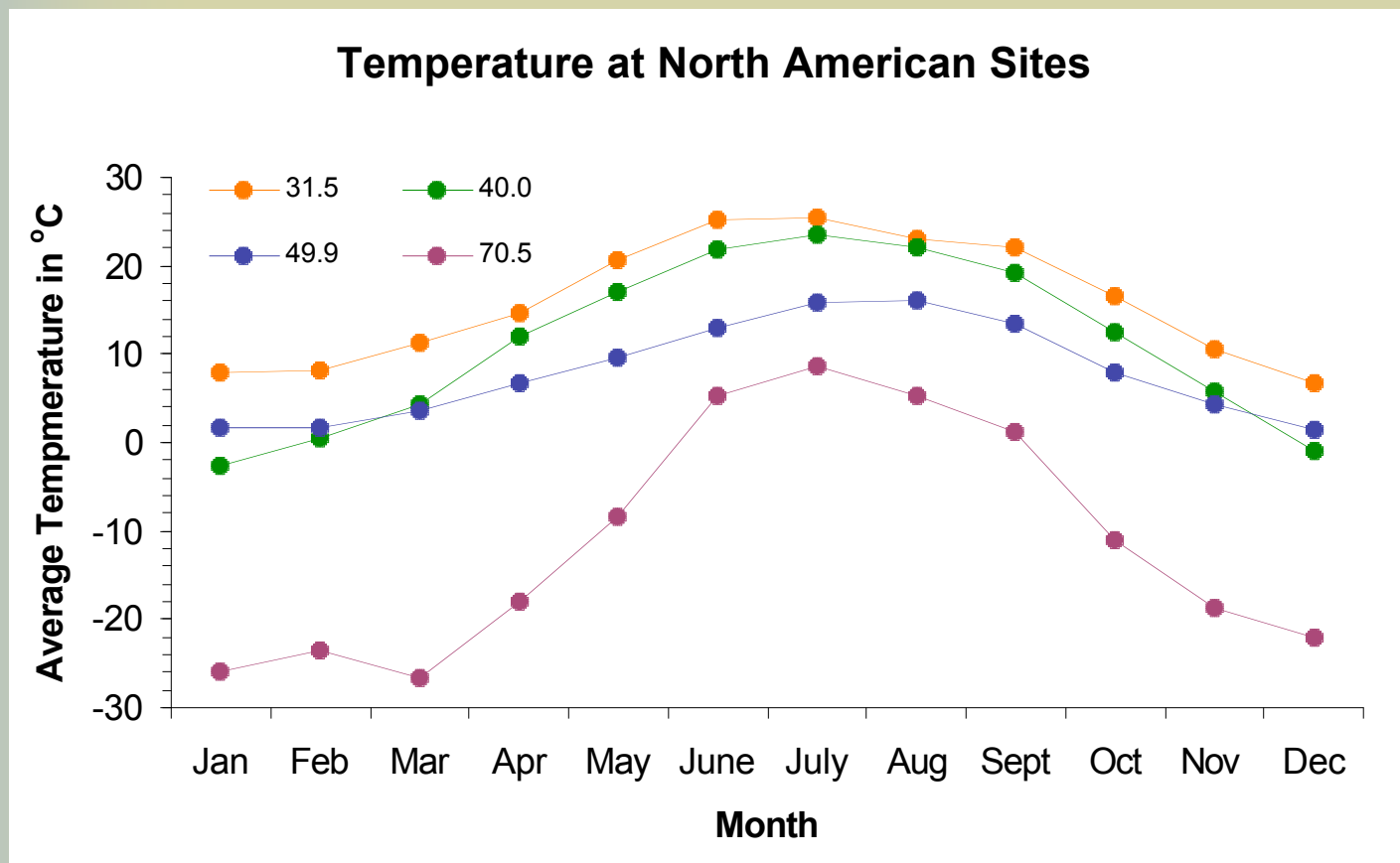
Other applications



***Plot created by Gretchen Miller of UC Berkeley**



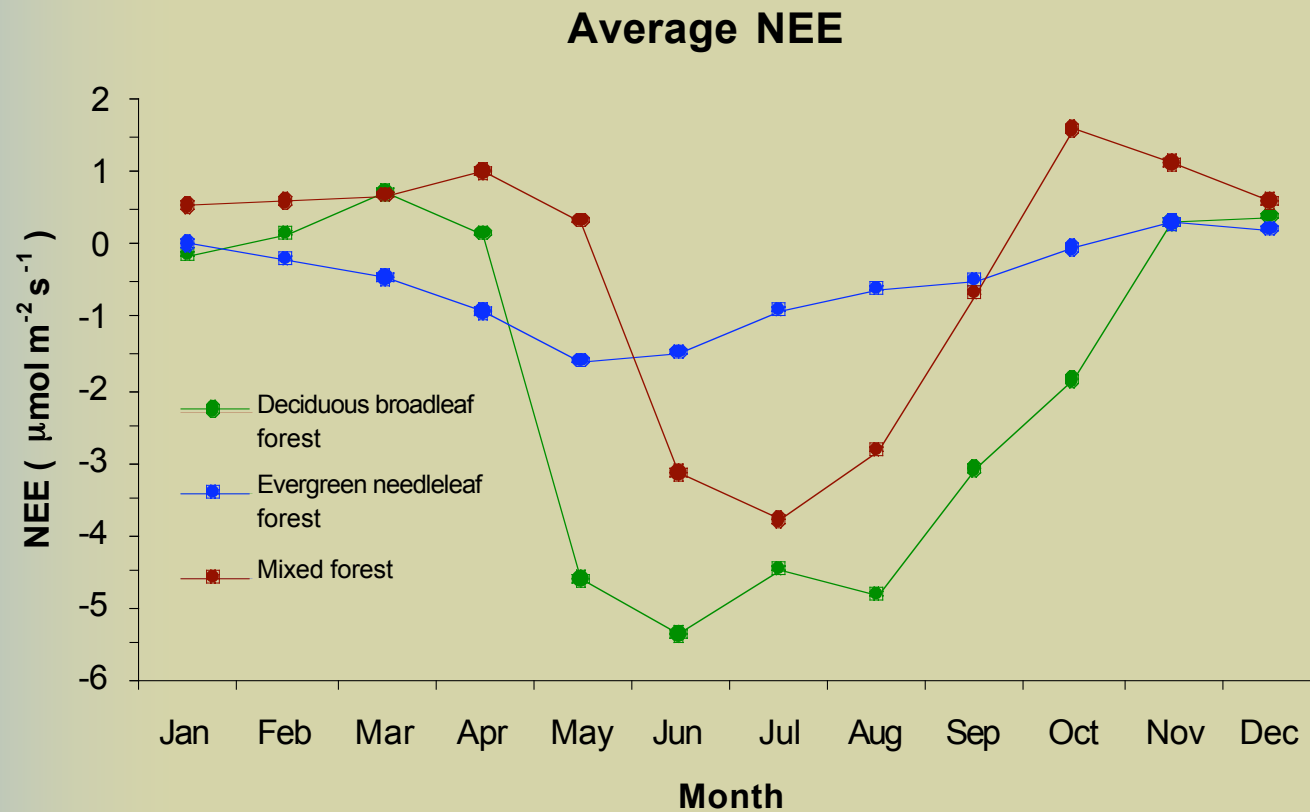
Observations by latitude



***Plot created by Gretchen Miller of UC Berkeley**



Observations by ecosystem type



*Plot created by Gretchen Miller of UC Berkeley



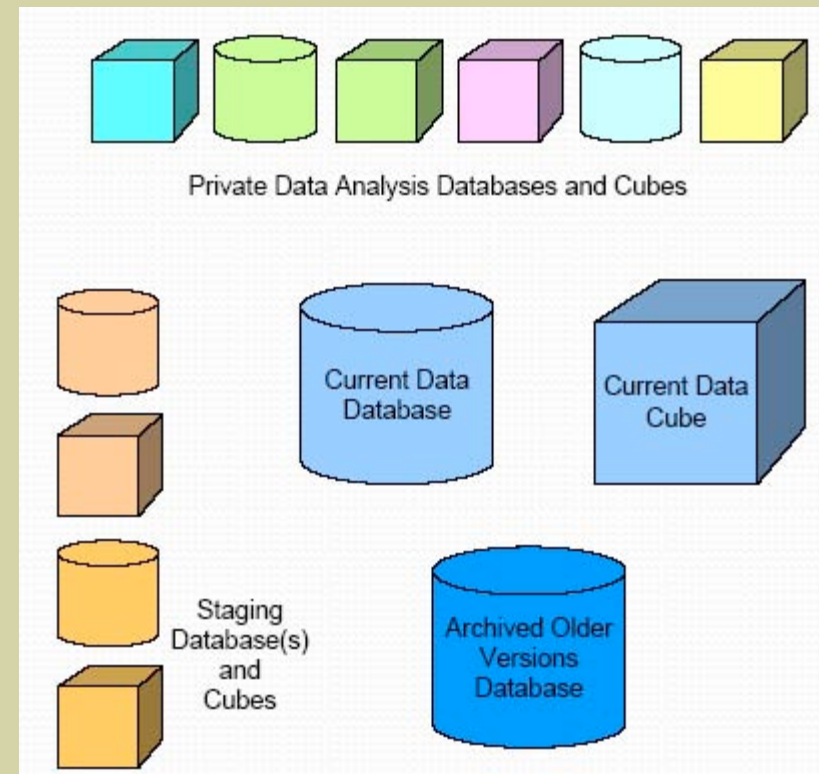
Some Lessons Learned so Far

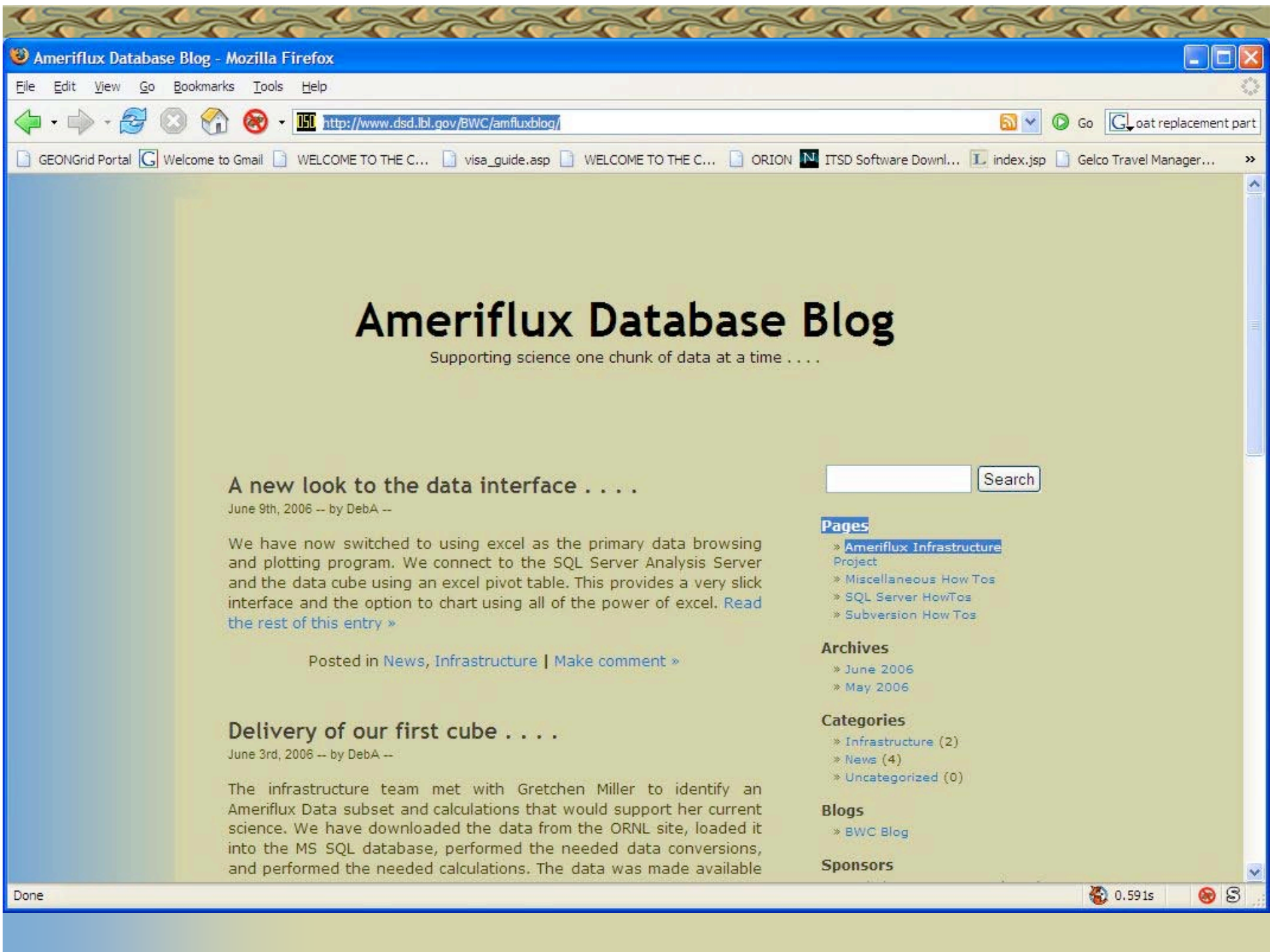
- Data naming and unit consistency is critical to easy ingest of large amounts of data
- Commercial tools do not necessarily provide all the right analysis capabilities directly
- Scaling capabilities of the tools not yet clear
- We will need tools to aid in notification of PIs



Portal Deployment

- Behind the portal are a collection of databases and data cubes
- Distribution for ease of use
 - Only see the data of interest
 - Private data remains stable
- Distribution for scaling
 - Smaller queries on smaller databases take less resources
 - Larger databases and cubes can be replicated across machines
- Batch job like infrastructure for managing very long running queries







Acknowledgements

- Science Team
 - Dennis Baldocchi
 - Bev Law
 - Gretchen Miller
- Cyberinfrastructure
 - Matt Rodriguez
 - Monte Goode
- Microsoft
 - Tony Hey
 - Nolan Li
- Oak Ridge National Lab CDIAC personnel
- Berkeley Water Center
 - Yoram Rubin
 - Susan Hubbard



URLs and Connection Coordinates

- Web Site

- <http://esd.lbl.gov/BWC>

- Blog

- <http://dsd.lbl.gov/BWC/amfluxblog>

- E-mail

- bwc-tci@lists.berkeley.edu




Berkeley Water Center - Mozilla

File Edit View Go Bookmarks Tools Window Help

Back Forward Reload Stop <http://esd.lbl.gov/BWC/> Search Print

Home Bookmarks

BERKELEY WATER CENTER



Effective water management is not purely a scientific problem, a political problem, a technological problem, a computer science problem nor a socioeconomic problem; it is a complex, 21st Century problem that demands collaborative coordination between all of these disciplines. The Berkeley Water Center has been developed to integrate expertise across disciplines in support of a new research mode for water investigations. The mission of the Berkeley Water Center is to:

- Develop a seamless integration of [LBNL](#) and [UCB](#) expertise and apply the expertise to water problems;
- Develop [Research Thrust Areas](#) (RTAs) that integrate Berkeley water expertise within those areas;
- Create collaborative [opportunities](#) between Berkeley Water Center and other expert groups and resources;
- Create strong, mutually beneficial [partnerships](#) between Berkeley and other academic, governmental, and private sector institutions.
- Accelerate development of RTA research results into applications through strategic [partnerships](#);
- Function as a [CITRIS member institution](#) and as the water arm of [BIE](#).

[OUR MISSION](#)

[MEMBERSHIP & GOVERNANCE](#)

[RESEARCH PARTNERS](#)

[RESEARCH THRUST AREAS](#)

[OPPORTUNITIES](#)

[ACTIVITIES](#)

[LOCATION & SURROUNDINGS](#)

[CONTACT US](#)

<http://esd.lbl.gov/BWC/>

